OpenEarthMap: A Benchmark Dataset for Global High-Resolution Land Cover Mapping

Junshi Xia¹,*, Naoto Yokoya²,1,*,†, Bruno Adriano¹,*, and Clifford Broni-Bediako¹

¹RIKEN AIP, Japan {junshi.xia,bruno.adriano,clifford.broni-bediako}@riken.jp
²The University of Tokyo, Japan yokoya@k.u-tokyo.ac.jp

Figure 1: A world map showing the locations of 97 regions included in OpenEarthMap and eight annotated examples.

Abstract

We introduce OpenEarthMap, a benchmark dataset, for global high-resolution land cover mapping. OpenEarthMap consists of 2.2 million segments of 5000 aerial and satellite images covering 97 regions from 44 countries across 6 continents, with manually annotated 8-class land cover labels at a 0.25–0.5m ground sampling distance. Semantic segmentation models trained on the OpenEarthMap generalize worldwide and can be used as off-the-shelf models in a variety of applications. We evaluate the performance of state-of-the-art methods for unsupervised domain adaptation and present challenging problem settings suitable for further technical development. We also investigate lightweight models using automated neural architecture search for limited computational resources and fast mapping. The dataset is available at https://open-earth-map.org.

1. Introduction

Land cover classification maps are the basic information for decision making in various applications, such as land use planning, food security, resource management, and disaster response. Meter-level resolution satellite imagery have been used to map the world, as represented by GlobeLand30 [6], FROM-GLC [5], and recent benchmarks such as OpenSentinelMap [20] and DynamicEarthNet [44]. Satellite imagery at a sub-meter level of ground sampling distance (GSD) enables the extraction of core map information such as buildings and roads. In recent years, there has been substantial progress in automatic construction of building footprints over large areas [39].

Since the advent of deep learning [2], a great deal of effort has been devoted to developing benchmark datasets for high-resolution remote sensing image analysis to facilitate advances in theory and practice. SpaceNet [47] and IEEE GRSS DFC [18], among others, regularly introduce benchmark datasets to the public through competitions that drive research and development. Building detection, road detection, object detection, and land cover classification (semantic segmentation) are the most typical tasks for which these datasets are used in supervised learning [57, 29]. Apart from supervised learning, these datasets have been used in more realistic problems, including transfer learning [49], semi-supervised learning [4] and weakly supervised learning [36, 21]. Benchmark datasets that contribute to solving social problems regarding change detection and disaster damage mapping have been developed as well [14, 16].

Benchmark datasets for semantic segmentation at sub-meter level resolution have two problems: regional disparity and annotation quality. The regions included in many benchmarks are often biased toward developed countries. Thus, benchmark datasets for regions where map information is not well maintained are scarce. Two main reasons why this problem has not been easily solved are the lack of high-resolution open aerial imagery in developing countries and that commercial high-resolution satellite imagery are basically not redistributable. Other than buildings and roads, the annotation quality of land cover labeling in existing benchmarks is coarse, even though images are at sub-meter level resolution. This is due to the high cost of manually labeling sub-meter-resolution imagery in spatial detail. Thus, most of the labeling data are based on OpenStreetMap [35] and open map data from local governments.

In this work, we propose OpenEarthMap, a benchmark dataset for global high-resolution land cover mapping with the goal of providing automated mapping for everyone. OpenEarthMap presents a major advance over existing data.
with respect to geographic diversity and annotation quality (see Table 1). OpenEarthMap consists of 8-class land cover labels at a 0.25–0.5m GSD of 5000 images, covering 97 regions from 44 countries across 6 continents. We adopted RGB images of some existing benchmark datasets for building detection and collected additional images for areas not covered by these benchmarks to balance the regional disparities. All images were manually labeled to ensure high-quality annotation. We evaluate the performance of state-of-the-art methods for semantic segmentation and unsupervised domain adaptation tasks and identify problem settings suitable for further technical development. In addition, lightweight models based on automated neural architectural search are investigated for cases where people requiring automated mapping have limited computational resources or for rapid mapping applications such as disaster response.

2. The Dataset

2.1. Source of Imagery

Our strategy is to reuse images from existing benchmark datasets as much as possible and manually annotate new land cover labels. We selected xBD [16], Inria [30], Open Cities AI [33], SpaceNet [47], Landcover.ai [3], AIRS [6], GeoNRW [1], and HTCD [35] datasets based on the condition that the source images are redistributable, the ground sampling distance (GSD) is equal to or less than 0.5m, and the images have geocoorordinate information. If there are enough images of a region, which we defined at a scale of province or city, we sampled 50–70 images of that region at a size of 1024 × 1024 pixels. The number of images from each dataset we adopted was determined based on the diversity and balance of the continents and countries where the images were taken. For countries and regions not covered by the existing datasets, aerial images publicly available in such countries or regions were collected to mitigate the regional gap, which is an issue in most of the existing benchmark datasets. The open data were downloaded from OpenAerialMap [34] and geospatial agencies [15, 32]. See the supplementary for more details of attribution.

In addition to this geographic diversity, our dataset includes a mixture of images taken from different platforms, including satellite, aircraft, and UAV. For very high-resolution images with GSD less than 0.25m, we resampled the images to 0.3m or 0.5m to account for object size and visual interpretability of the captured area. Basically, the images were selected by a combination of random sampling and manual checking for each region. Moreover, if the number of images of a particular region is very large in the source benchmark dataset, we trained a segmentation model using sequentially labeled data (e.g., every 10 images) and another regression model to estimate the loss. Then, we added the images that have high values of predicted loss, as they are more difficult by a model trained with the available labels to segment.

In the end, we collected a total of 5000 images from 97 regions of six continents. Figure 1 shows annotated samples and the geographic distribution of the 97 regions with different colors indicating the source datasets. Figure 2 depicts the number of images in our dataset for each of the six continents, colored to indicate the origin of the images. Asia, Africa, and South America are not well covered by the source datasets; thus, we added many images from public data to balance the regional disparities. Figure 3 presents a t-SNE 2D plot based on the similarity of image features for the 97 regions. For each region, we used the average of features extracted by EfficientNet-B4 trained as an encoder of U-Net on OpenEarthMap. The 12 representative images in the 2D plot show that different locations correspond to diverse images. It can also be seen that the different source datasets are complementary to each other, and that the diversity of images is enriched by the open data we added. The different symbols correspond to the six continents and enable the similarities between the continents to be seen. For example, regions in Europe and North America as well as Africa and South America are similar in the image features.

2.2. Classes, Annotations, and Data Split

Classes: We provide annotations with eight classes: bareland, rangeland, developed space, road, tree, water, agriculture land, and building. The class selection is consistent with existing products and benchmark datasets (e.g., LoveDA [49] and DeepGlobe [12]) with sub-meter GSD.

Table 1: Summary of remote sensing benchmark datasets for semantic segmentation. B: building extraction, R: road extraction, LC: land cover mapping, and CD: change detection. The number of segments was counted on available labels.

<table>
<thead>
<tr>
<th>Image level</th>
<th>GSD (m)</th>
<th>Dataset</th>
<th>Task</th>
<th>Classes</th>
<th>Countries</th>
<th>Regions</th>
<th>Area (km2)</th>
<th>Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meter level</td>
<td>10</td>
<td>OpenSentinelMap [20]</td>
<td>LC</td>
<td>15</td>
<td>—</td>
<td>—</td>
<td>505,202</td>
<td>3,467,352</td>
</tr>
<tr>
<td>3</td>
<td>DynamicEarthNet [44]</td>
<td>LC/CD</td>
<td>7</td>
<td>—</td>
<td>75</td>
<td>707</td>
<td>897,855</td>
<td></td>
</tr>
<tr>
<td>Sub-meter level</td>
<td>0.3–0.5</td>
<td>SpaceNet 1&amp;2 [47]</td>
<td>B</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5,555</td>
<td>685,235</td>
</tr>
<tr>
<td>0.5/0.3/0.5</td>
<td>DeepGlobe [12]</td>
<td>R/B/LC</td>
<td>2/2/7</td>
<td>—</td>
<td>—</td>
<td>2,220/984/1,717</td>
<td>—/302,701/20,697</td>
<td></td>
</tr>
<tr>
<td>0.02–0.2</td>
<td>Open Cities AI [33]</td>
<td>B</td>
<td>2</td>
<td>8</td>
<td>11</td>
<td>419</td>
<td>792,484</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>xBD [16]</td>
<td>B/C/D</td>
<td>2/4</td>
<td>15</td>
<td>21</td>
<td>3,382</td>
<td>850,736</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>LoveDA [49]</td>
<td>LC</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>536</td>
<td>166,768</td>
<td></td>
</tr>
<tr>
<td>0.25–0.5</td>
<td>OpenEarthMap</td>
<td>LC</td>
<td>8</td>
<td>44</td>
<td>97</td>
<td>799</td>
<td>2,205,395</td>
<td></td>
</tr>
</tbody>
</table>
Table 2: The number and proportion of pixels and the number of segments of the eight classes.

<table>
<thead>
<tr>
<th>Color (HEX)</th>
<th>Class</th>
<th>Pixels</th>
<th>Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>800000</td>
<td>Bareland</td>
<td>74</td>
<td>1.5</td>
</tr>
<tr>
<td>00FF24</td>
<td>Rangeland</td>
<td>1130</td>
<td>22.9</td>
</tr>
<tr>
<td>949494</td>
<td>Developed space</td>
<td>798</td>
<td>16.1</td>
</tr>
<tr>
<td>FFFFFF</td>
<td>Road</td>
<td>331</td>
<td>6.7</td>
</tr>
<tr>
<td>226126</td>
<td>Tree</td>
<td>996</td>
<td>20.2</td>
</tr>
<tr>
<td>0045FF</td>
<td>Water</td>
<td>161</td>
<td>3.3</td>
</tr>
<tr>
<td>226126</td>
<td>Agriculture land</td>
<td>680</td>
<td>13.7</td>
</tr>
<tr>
<td>DE1F07</td>
<td>Building</td>
<td>770</td>
<td>15.6</td>
</tr>
</tbody>
</table>

Figure 2: The number of images of the six continents in OpenEarthMap.

Figure 3: t-SNE 2D visualization of the 97 regions based on features extracted with EfficientNet-B4 trained on OpenEarthMap. The images are samples of 12 regions with horizontal bar charts of class proportions attached at the bottom.

Table 2 shows the number and proportion of labeled pixels and the number of segments of each class. Here, as well as in Table 1, we refer to a segment as a set of connected pixels with the same label, and it was counted using OpenCV’s findContours function. It can be seen that the elevated objects (e.g., tree and building) are finely annotated compared to the ground objects (e.g., agriculture land). As can be seen in the horizontal bar charts of 12 representative regions in Figure 3, the class proportions in the different regions are diverse.

**Annotations:** A total of 16 people worked on the annotation process: 8 people were responsible for annotating the images, while the remaining 8 people performed quality checks to point out errors. One person labels an image and at least two people perform the quality check. We spent a longer time labeling the first 100 images and exchanging ideas with each other to ensure that all participants were in agreement about the class definitions. On average, the labeling took 2.5 hours per image. This is significantly longer than the 1.5 hours of Cityscapes [11], which illustrates the difficulty of labeling remote sensing images. All the labeling was done manually. For the labeling of images of the existing benchmark datasets, only the building class was used as the starting point. However, since a lot of label noise was found, segments of buildings were also manually modified. The most important feature of OpenEarthMap’s labeling is its level of spatial detail. As shown in Table 1, the area covered by the images in OpenEarthMap is not very large compared to the other benchmark datasets, however, the number of segments is 10 times more than that of LoveDA.

The accuracy of human annotations is evaluated by having two different people labeling 200 images twice. We selected two or three images with as many classes as possible based on the first annotation from each region to constitute the 200 images. The percentage of pixels that were labeled as the same class in the two different annotations by different people is 78%. This percentage is significantly lower than the 96% in Cityscapes [11], suggesting that annotation of high-resolution remote sensing images is much more challenging than annotation of urban street scenes. The relationship between human labeling accuracy and estimation accuracy of the state-of-the-art segmentation models is discussed in Section 3.4.

**Data split:** For the semantic segmentation task, the images from each region were randomly divided into training, validation, and test sets with a ratio of 6:1:3, which respectively yielded 3000, 500, and 1500 images out of the total 5000 images. To ensure that all classes in each region are included in the training set and as many classes as possible are included in the test set, the split with the least mismatch between the training and test classes was selected from multiple random trials. For the unsupervised domain adaptation (UDA) tasks, we adopt two ways of data split to investigate regional-level and continent-wise domain gaps. For regional-level UDA, the entire dataset is divided into 73 and 24 regions for source and target domains, respectively. The split was performed in such a way that both the source and target domains consist of relatively even distribution of the countries from all six continents as well as a balance between urban and rural areas. This split is not as extreme as the urban-rural split in LoveDA but rather it is a realistic scenario in domain adaptation where OpenEarthMap is at hand as source data and adapts models for mapping in any new region, not only urban-rural adaptation. For continent-wise UDA, we use data from one continent as the source domain and other continents as the target domains.

**2.3. Comparison with Related Datasets**

Very recently, meter-level resolution benchmarks have made great progress in global land cover mapping;
OpenSentinelMap [20] is featured in its comprehensive coverage of the globe exploiting open data of Sentinel-2 and OpenStreetMap while DynamicEarthNet [44] is advantageous at high-temporal resolution. OpenEarthMap goes one step further in providing spatially detailed annotation at the sub-meter level. A more detailed comparison is made with LoveDA [49] and DeepGlobe [12], which have similar resolution and class definitions as OpenEarthMap. Figure 4(a) shows a comparison of the class proportions of the three datasets. It should be noted that LoveDA does not include rangeland, and that in the DeepGlobe dataset for land cover classification, buildings and roads are included in the urban class. There is no dominant class in OpenEarthMap and the class proportions are relatively balanced. The normalized histogram of the number of segments in a single image is shown in Figure 4(b). In terms of image size, LoveDA is the same (1024×1024 pixels) as OpenEarthMap, while DeepGlobe is larger (2448×2448 pixels). The histogram of OpenEarthMap has a very long tail, showing a much larger number of segments in each image of OpenEarthMap than the other datasets. The spatially detailed labeling of the OpenEarthMap is reflected in the cross-dataset evaluation and the out-of-sample prediction results of trained models presented in Sections 5 and 6.

3. Land Cover Semantic Segmentation

3.1. Baselines

For the land cover semantic segmentation task, CNN-based and Transformer-based architectures were evaluated and compared on the OpenEarthMap dataset. More specifically, the chosen models are U-Net [27], U-NetFormer [50], FT-U-NetFormer [50], DeepLabV3 [7], HRNet [41], SETR [56], SegFormer [54], and UPerNet [53] with backbones of ViT [13], Twins [10], Swin Transformer [26], ConvNeXt [26], and K-Net [55].

3.2. Results

General results: The results obtained on the test set of OpenEarthMap are presented in Table 3. The main findings are discussed as follows: (1) U-Net with EfficientNet-B4 as backbone outperforms both U-Net with ResNet-34 and U-Net with VGG-11. The reason might be that EfficientNet-B4 is more effective for extracting relevant features, and to that effect, both high-level features and low-level spatial information are used for robust segmentation. (2) UPerNet with Swin-B and Twins, as well as SegFormer and K-Net perform better than DeepLabV3 and HRNet. This might be attributed to the strong modeling capabilities and dynamic feature aggregation of Swin-B, Twins, and MiT-B5. (3) U-NetFormer and FT-U-NetFormer share the top positions because both methods adopt a global–local Transformer block to construct global and local information in the decoder, and use advanced encoder (e.g., ResNeXt and Swin-B) to extract features. (4) UPerNet with ViT and ConvNeXt, along with SETR obtain worse results than other Transformer-based models. Two reasons might be that the hyperparameters (e.g., optimizer and learning rate) of these methods may need to be carefully tuned, and advanced data augmentation may be required for transfer learning from ImageNet to the OpenEarthMap dataset. In all, considering performance along with the number of parameters and FLOPs, U-Net-EfficientNet-B4, UPerNet-Swin-B, and FT-U-NetFormer are recommended.

Visualization: Examples of segmentation results obtained from some selected methods are presented in Figure 5. The U-Net-EfficientNet-B4 and FT-U-NetFormer produce the best detailed visualization results. In the first row of Figure 5, DeepLabV3 wrongly classified the water area of the dam as rangeland while other methods identified them. In the second row, U-Net-EfficientNet-B4, SegFormer and FT-U-NetFormer were able to identify the tiny roads in the top-right parts of the image. Water and bareland classes respectively achieved the highest and the lowest accuracies in all methods. The boundaries of the buildings and the roads were difficult to identify properly because of disorganized layouts and varying sizes. Rangeland, agricultural land and trees are easy to confuse due to the similarities in their spectra. Roads were easily misclassified as developed space because parking lots and cover materials in some rural areas are quite similar.

3.3. Neural Architecture Search

LoveDA [49], DeepGlobe [12], and other previous benchmarks [3, 30, 8] were experimented with only manually designed networks [17, 41, 7, 27, 23] for the semantic segmentation task. In contrast, we further experimented the OpenEarthMap dataset with two automated neural architecture search methods, SparseMask [52] and FasterSeg [9], by automatically searching for compact segmentation architectures. Such architectures might offer a useful baseline for research in the field of automated neural architecture search in remote sensing with OpenEarthMap. Following the ar-
Table 3: Semantic segmentation results of the baseline models on the test set of the OpenEarthMap dataset. The results are based on test-time augmentation (TTA), in particular flipping.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Bareland</th>
<th>Rangeland</th>
<th>Developed</th>
<th>Road</th>
<th>Tree</th>
<th>Water</th>
<th>Agriculture</th>
<th>Building</th>
<th>IoU (%)</th>
<th>mIoU (%)</th>
<th>Params (M)</th>
<th>FLOPs (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>VGG-11</td>
<td>40.69</td>
<td>56.76</td>
<td>53.99</td>
<td>62.16</td>
<td>72.44</td>
<td>82.81</td>
<td>73.14</td>
<td>77.77</td>
<td>64.97</td>
<td>18.26</td>
<td>233.33</td>
<td></td>
</tr>
<tr>
<td>U-Net</td>
<td>ResNet-34</td>
<td>40.35</td>
<td>57.75</td>
<td>54.92</td>
<td>62.87</td>
<td>72.65</td>
<td>82.24</td>
<td>74.06</td>
<td>78.58</td>
<td>65.43</td>
<td>24.44</td>
<td>126.68</td>
<td></td>
</tr>
<tr>
<td>U-Net</td>
<td>EfficientNet-B4</td>
<td>50.63</td>
<td>58.17</td>
<td>56.27</td>
<td>64.83</td>
<td>73.20</td>
<td>86.02</td>
<td>76.28</td>
<td>80.20</td>
<td>68.20</td>
<td>20.30</td>
<td>45.47</td>
<td></td>
</tr>
<tr>
<td>U-NetFormer</td>
<td>ResNeXt101</td>
<td>46.09</td>
<td>60.67</td>
<td>58.12</td>
<td>65.07</td>
<td>73.77</td>
<td>86.34</td>
<td>76.98</td>
<td>79.96</td>
<td>68.37</td>
<td>19.21</td>
<td>76.92</td>
<td></td>
</tr>
<tr>
<td>FT-U-NetFormer</td>
<td>Swin-B</td>
<td>50.19</td>
<td>60.84</td>
<td>57.58</td>
<td>65.85</td>
<td>73.33</td>
<td>87.44</td>
<td>77.50</td>
<td>80.29</td>
<td>69.13</td>
<td>20.30</td>
<td>49.98</td>
<td></td>
</tr>
<tr>
<td>DeepLabV3</td>
<td>ResNet-50</td>
<td>39.11</td>
<td>56.16</td>
<td>52.28</td>
<td>60.57</td>
<td>71.25</td>
<td>79.32</td>
<td>70.75</td>
<td>75.83</td>
<td>63.16</td>
<td>68.14</td>
<td>269.76</td>
<td></td>
</tr>
<tr>
<td>HRNet</td>
<td>W48</td>
<td>39.71</td>
<td>55.50</td>
<td>53.49</td>
<td>59.22</td>
<td>71.10</td>
<td>79.03</td>
<td>71.38</td>
<td>75.12</td>
<td>63.07</td>
<td>59.46</td>
<td>236.68</td>
<td></td>
</tr>
<tr>
<td>UPerNet</td>
<td>ViT</td>
<td>34.39</td>
<td>54.45</td>
<td>50.64</td>
<td>54.57</td>
<td>69.73</td>
<td>79.24</td>
<td>66.22</td>
<td>74.92</td>
<td>60.52</td>
<td>144.17</td>
<td>395.07</td>
<td></td>
</tr>
<tr>
<td>UPerNet</td>
<td>Swin-B</td>
<td>44.52</td>
<td>58.98</td>
<td>54.78</td>
<td>63.43</td>
<td>72.20</td>
<td>83.71</td>
<td>72.97</td>
<td>78.11</td>
<td>66.09</td>
<td>59.94</td>
<td>236.08</td>
<td></td>
</tr>
<tr>
<td>SegFormer</td>
<td>MiT-B5</td>
<td>36.84</td>
<td>57.94</td>
<td>53.53</td>
<td>63.60</td>
<td>70.51</td>
<td>80.11</td>
<td>72.21</td>
<td>77.35</td>
<td>64.01</td>
<td>81.97</td>
<td>51.86</td>
<td></td>
</tr>
<tr>
<td>SETr PUP</td>
<td>ViT-L</td>
<td>45.35</td>
<td>55.72</td>
<td>51.31</td>
<td>55.47</td>
<td>67.63</td>
<td>73.12</td>
<td>67.14</td>
<td>75.48</td>
<td>61.40</td>
<td>309.35</td>
<td>212.45</td>
<td></td>
</tr>
<tr>
<td>UPerNet</td>
<td>Twins</td>
<td>37.29</td>
<td>57.62</td>
<td>53.83</td>
<td>60.23</td>
<td>72.32</td>
<td>81.93</td>
<td>71.71</td>
<td>77.49</td>
<td>64.05</td>
<td>90.96</td>
<td>250.91</td>
<td></td>
</tr>
<tr>
<td>UPerNet</td>
<td>ConvNeXt</td>
<td>40.61</td>
<td>54.94</td>
<td>51.76</td>
<td>58.47</td>
<td>70.44</td>
<td>75.95</td>
<td>68.94</td>
<td>74.30</td>
<td>61.93</td>
<td>122.1</td>
<td>292.42</td>
<td></td>
</tr>
<tr>
<td>K-Net</td>
<td>Swin-B</td>
<td>44.02</td>
<td>57.81</td>
<td>54.85</td>
<td>62.91</td>
<td>71.76</td>
<td>85.18</td>
<td>73.41</td>
<td>78.91</td>
<td>66.11</td>
<td>246.97</td>
<td>419.51</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Visual comparison of land cover mapping results of some of the baseline models presented in Table 3.

Table 4: Lightweight models discovered on OpenEarthMap training set. FLOPs and FPS are measured on 1024 × 1024 input, and mIoU on the test set of OpenEarthMap.

<table>
<thead>
<tr>
<th>Method</th>
<th>Trial</th>
<th>Params (M)</th>
<th>FLOPs (G)</th>
<th>FPS (ms)</th>
<th>No TTA TTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SparseMask</td>
<td>1st</td>
<td>2.96</td>
<td>10.28</td>
<td>51.2</td>
<td>58.23</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>3.10</td>
<td>10.39</td>
<td>52.2</td>
<td>58.06</td>
</tr>
<tr>
<td>FasterSeg</td>
<td>1st</td>
<td>2.23</td>
<td>14.58</td>
<td>143.2</td>
<td>57.55</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>3.47</td>
<td>15.37</td>
<td>143.3</td>
<td>57.55</td>
</tr>
</tbody>
</table>

3.4. Human Annotation vs Machine Prediction

As mentioned in Section 2.2, 200 images were labeled twice by different people. The remaining 4800 images were used to train UPerNet with Swin-B to compare the quality of human labeling with the results from the machine. To effectively investigate the comparison, the number of training images was varied from 10% to 100%; the results are shown in Figure 6. It can be seen that with 50% of the training images, the machine attains almost the same level of human annotation and larger training percentages improve the accuracy (see Figure 6a). For human annotation, the challenging classes include bareland, rangeland, and tree. For bareland, rangeland, developed space, and tree classes, 50%, 30%, 50%, and 10% of the training set, respectively, yielded better results than the ones of human annotation (see Figure 6b). The challenging class for the machine is agriculture land, where it trails behind human annotation by 2.3%. Regarding road, water, and building classes, with 100% of the training images, the machine slightly (< 0.34%) trails behind the human annotation.
3.5. Learning from Limited Labels

We also investigated the performance of CNN-based (U-Net-EfficientNet-B4) and Transformer-based (SegFormer, UPerNet-Swin-B and K-Net) models on limited training samples. Table 5 presents the results of using only 10% of the OpenEarthMap training set to train the models. It is apparent from Table 5 that U-Net-EfficientNet-B4 outperforms all three Vision Transformer-based methods in all the class-specific IoUs by about 6-15%. The main reason is that the representation capacity of ViTs typically lacks the inductive bias in CNNs. Therefore, ViTs require more training data than CNNs. We believe that Vision Transformers with small-sized data or limited labels is an interesting topic that requires further study. Moreover, data augmentation, regularization, and tuning of hyper-parameters still need to be explored when training on limited training data.

4. Unsupervised Domain Adaptation

4.1. Baselines

For the unsupervised domain adaptation task, a metric-based method (MCD [46]), adversarial training methods including AdaptSeg [45], category-level adversarial network (CLAN) [28], TransNorm [51], and fine-grained adversarial learning framework for domain adaptive (FADA) [48], as well as self-training methods including pyramid curriculum DA (PyCDA) [22], class-balanced self-training (CBST) [58], instance adaptive self-training (IAST) [21], and DAFormer [17] are adopted. DAFormer is based on SegFormer and the others are based on DeepLabV2.

4.2. Results

Regional-level UDA: We investigated the regional-level domain gap since different regions in the same continent might suffer from a distribution shift. The results obtained on the test set of 24 regions of OpenEarthMap are presented in Table 6. In general, the Oracle settings obtained the best results. Due to the regional domain gap, the source-only settings yielded the lowest accuracy. The results of source-only SegFormer are significantly better than those of source-only DeepLabV2. Compared to manufactured classes (i.e., building and road), the accuracies of natural classes (i.e., water and bareland) decreased significantly. With the exception of TransNorm, the adversarial training methods did not perform well on this task due to the diversity in the OpenEarthMap dataset. TransNorm slightly improved the performance because the source and the target images have distinct spectral statistics since they were taken from different sensors and regions. The class imbalance problem is addressed using pseudo-label creation via the CBST and IAST techniques, resulting in higher performance. Due to better domain generalization of SegFormer and effective training strategy in self-training, DAFormer obtained the best mIoU of 62.35%. Visual examples of the UDA results are presented in Figure 7. In the first row of Figure 7, source-only DeepLabV2 can barely identify the water area (top-right) and the roads (bottom-right). IAST and CBST performance improves for water but they lose the ability to recognize the roads. DAFormer performs very well in the two complex areas. In the second row, DAFormer shows better visualization results in the small water area (top-right) and the boundaries of roads and buildings than the other UDA methods.

Continent-wise UDA: We also investigated the continent-wise domain gap on the OpenEarthMap dataset using U-Net-EfficientNet-B4, SegFormer, and DAFormer. The results are presented in Figure 8. Compared to the UDA settings (e.g., GTA5→Cityscapes with similar content and different style) in computer vision and previous settings in remote sensing (e.g., urban→rural in LoveDA), UDA on continent-wise has larger content and style gaps. The lim-
extension of DAFormer or new UDA method with U-Net-EfficientNet-B4.

5. Cross-Dataset Evaluation

In this section, we evaluate the advantage of using the OpenEarthMap dataset as a starting point (fine-tuning) in the semantic segmentation task over other open-source land cover mapping datasets. Here we compare OpenEarthMap with LoveDA [49] and DeepGlobe [12]. We adopted the same U-Net model with an EfficientNet-B4 as a backbone listed in Table 3 and trained from scratch on the three datasets using similar training settings as Section 3.2. Then, we fine-tuned the OpenEarthMap and the LoveDA pre-trained models on the DeepGlobe dataset. Similarly, the OpenEarthMap and the DeepGlobe pre-trained models were fine-tuned on the LoveDA dataset. All the experiments were run threefold; we report the mean and the standard deviation segmentation accuracy for 20 epochs. As presented in Figure 9, the results indicate that using a model that is pre-trained on OpenEarthMap as a starting point could yield better performance than models pre-trained on LoveDA and DeepGlobe. For example, when fine-tuned on the DeepGlobe dataset, the initial IoU score of the OpenEarthMap pre-trained model is about 4% higher than the fully-trained model on DeepGlobe (see Figure 9a). Although the OpenEarthMap pre-trained model is slightly lower than the LoveDA pre-trained one in early epochs, OpenEarth-
Chesapeake Bay land cover map

Other

Bareland

LoveDA, we adopted six common classes (produced by U-Net models trained on OpenEarthMap and results of the Chesapeake Bay land cover tool with those map consists of 13 classes. To fairly compare the mapping to evaluate performance. The Chesapeake Bay land cover map [42] was used as reference.

To further investigate the generalization performance of a model trained on OpenEarthMap, we created land cover classification maps from out-of-sample imagery (i.e., images that are not included in OpenEarthMap). The NAIP image is the source data.

Map increasingly outperforms both models as the number of epochs increases. Furthermore, when fine-tuned on the LoveDA dataset, the OpenEarthMap pre-trained model attains a more than 20% increase in the initial IoU score, and its performance remains higher when the number of epochs increases (see Figure 9b).

6. Demonstration on Out-of-Sample Imagery

To further investigate the generalization performance of a model trained on OpenEarthMap, we created land cover classification maps from out-of-sample imagery (i.e., images that are not included in OpenEarthMap). See the supplementary for more results. Here we present a map created from an NAIP [43] image resampled at 0.5m GSD. A Chesapeake Bay land cover map [42] was used as reference to evaluate performance. The Chesapeake Bay land cover map consists of 13 classes. To fairly compare the mapping results of the Chesapeake Bay land cover tool with those produced by U-Net models trained on OpenEarthMap and LoveDA, we adopted six common classes (bareland, other, road, tree, water, and building) among the datasets and performed quantitative evaluation. Table 7 shows the accuracy of the land cover mapping for an area of approximately 15km × 28km in US, spanning from Washington, DC to Maryland. The IoUs from OpenEarthMap model are significantly higher than those from LoveDA, and the scores are high enough for practical mapping except bareland. The accuracy of bareland is low due to inconsistency in class definitions. For example, in the Chesapeake Bay land cover map, a construction site is labeled as bareland, while OpenEarthMap labels the same area as developed space. Figure 10 shows a visual example of the mapping results. Note that unlike the quantitative evaluation in Table 7, the vegetation classes (low vegetation, agriculture land, and range-land) that differ among the datasets are visualized in different colors. The OpenEarthMap model result is similar to the Chesapeake Bay land cover map in both classification and resolution, and achieved very fine spatial segmentation compared to the LoveDA model. This demonstrates the advantage of OpenEarthMap over LoveDA and how finely OpenEarthMap’s annotations are spatially detailed.

7. Conclusion and Societal Impacts

The existing benchmarks for land cover classification at sub-meter resolution lack regional diversity and annotation quality. To address this problem, we introduce OpenEarthMap, a benchmark dataset, for global high-resolution land cover mapping. The diversity of the dataset is shown in the coverage of 97 regions from 44 countries across 6 continents, while its finely detailed annotations are reflected in the generalization of the feature space. To demonstrate the practical usefulness of OpenEarthMap, we perform baseline experiments with several state-of-the-art models for semantic segmentation and UDA tasks, and create land cover maps for out-of-sample imagery to show that models trained on OpenEarthMap can adapt and generalize across the globe. We also demonstrate the challenges of the continent-wise domain gap and limited data training. We experiment NAS-based lightweight models for mapping with resource-limited devices. Further technical development is needed to improve the performance in continent-wise domain adaptation, limited training data, and lightweight models on OpenEarthMap for worldwide evaluation. The dataset is made publicly available for other researchers to build on it and create new practical tasks.

Societal Impacts: OpenEarthMap models could enable automated mapping of any location on Earth, which can support decision making in disaster response, environmental conservation, and urban planning. However, such models will make it easy for anyone to access map information related to national security as well as privacy if sub-meter resolution images are available. Appropriate data analysis ethics and data policies are required to avoid security and privacy breaches.

Acknowledgement

This work was supported by JST FOREST Grant Number JPMJFR206S, Japan.
References


[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:


